# *Register variation in spoken British English: the case of verb-forming suffixation*

This is a pre-publication version accepted by the International Journal of Corpus Linguistics, to appear in 2018. Please refer to the published version if you wish to quote from it.

# Register variation in spoken British English

## The case of verb-forming suffixation

Jacqueline Laws and Chris Ryder

University of Reading

The aim of this paper is to identify the effect of register variation in spoken British English on the occurrence of the four principal verb-forming suffixes: -*ate*, -*en*, -*ify* and -*ize*, by building on the work of Biber et al. (1999), Plag et al. (1999) and Schmid (2011). Register variation effects were compared between the less formal Demographically-Sampled and the more formal Context-Governed components of the original 1994 version of the British National Corpus. The pattern of -*ize* derivatives revealed the most marked register-based differences with respect to frequency counts and the creation of neologisms, whereas -*en* derivatives varied the least compared with the other three suffixes. Quantitative and qualitative analyses of these suffix profiles in the context of spoken language reveal markers of register formality that have not hitherto been explored; derivative usage patterns provide an additional dimension to previous research on register variation which has mainly focused on grammatical and lexical features of language.

**Keywords:** register variation, derivational morphology, spoken language, verb-forming suffixation

## 1. Introduction

It has been well-established in the literature on register variation (Biber 1988, Biber et al. 1999) that more formal written contexts, such as academic prose, require different linguistic structures compared with less formal contexts, such as fiction and conversation. The spoken register has similarly been extensively researched and the characteristics of various sub-registers (face-to-face and phone conversations, debates, interviews, broadcasts, spontaneous and planned speeches) have been successfully profiled, together with a host of written registers ranging from personal letters to academic prose (Biber 1988). The majority of research based on Biber's (1988) multidimensional analysis of register-based linguistic features has focused mainly on grammatical, lexical and discourse aspects of spoken and written language, e.g. Friginal et al. (2013) and Mazgutova & Kormos (2015), to name just two; whereas very little attention has been given to the role of derived forms as a function of register. The aim of the current study is to identify the relationship between usage patterns of complex verb forms and register formality in spoken British English.

## 2. Register variation and complex words

Biber's (1988) seminal and extensively cited work only identifies nominalisation, through the process of derivation, as one of the key linguistic features that varies most markedly between spoken and written registers. A few studies have subsequently compared the distributional characteristics of a range of derived forms between a single spoken corpus (conversation) and a variety of written corpora: prefixes and suffixes forming nouns, verbs, adjectives and adverbs (Biber et al. 1999, Schmid 2011), and noun-forming suffixes (Guz 2009; Säily 2011, Säily & Suomela 2017). These register-based studies reveal that the frequency of derivatives varies considerably between spoken and written registers, as well as within the variety of written domains; the key finding being that the range of complex lexemes (types) and their representation in a corpus (tokens) increases as the formality of the register increases. For example, Schmid (2011: 152) reports an increase in the proportion of prefixed words (tokens) occurring in conversation from 4.81% to 9.28% in letters and then to 43.30% in academic prose. Of the suffix categories analysed, Schmid (2011: 181) finds that the four principal verb-forming suffixes, *-ate*, *-en*, *-ify* and *-ize*, increased the most dramatically across the 5 registers, from 3% in conversation to 10% in letters and to 58% in academic prose, the last of these being the highest percentage produced by all the suffix categories in that study.

The work of Biber et al. (1999) and Schmid (2011) has provided clear empirical evidence that derived forms are required for expressing information transparently and succinctly. However, to date, with the exception of a few studies discussed below (Plag et al. 1999, Cowie 2006), the role of morphology as a marker of formality in speech and the effect of different spoken sub-registers on the usage patterns of all categories of derived words has not been systematically explored.

It is this gap that the current study addresses by focusing on the usage patterns of the four principal verb-forming suffixes in British English as a function of spoken register. The two sub-components of the original 1994 version of the British National Corpus (BNC) provide an ideal opportunity to examine such an effect: the distributional characteristics of derivatives in the less formal Demographically-Sampled (DS) component of the BNC can be compared with their corresponding profiles in the more formal Context-Governed (CG) component. As mentioned above, only a few studies have explored the effect of context formality on the usage patterns of derivatives between different speech registers. Plag et al. (1999) compare the whole written component of the BNC with the DS and CG sub-component spoken registers in relation to 15 suffixes (6 noun-forming, 1 verb-forming and 8 adjective-forming) and Cowie (2006) examines the characteristics of the suffix *-wise* across these three components of the BNC. These studies focus on the measurement of the productivity of certain suffix categories, where productivity was defined as the ability of an affix category to create new members; an overview of these metrics and their relevance to the current study are given in Section 3.

Plag et al. (1999) obtain systematic differences between the two spoken and the written components of the BNC, but of particular interest here are the effects found between the two speech registers: the analyses clearly indicate that productivity patterns in the more formal spoken register (CG) more closely resemble those observed in the written corpus than the less formal context (DS). In other words, the positive correlation observed between the frequency of occurrence of derived forms and register formality with respect to spoken and written language (Biber et al. 1999, Schmid 2011) was also observable across levels of speech formality. In light of these findings, the aim of the current study is to provide a systematic profile of the register-based characteristics of the four principal verb-forming suffixes in spoken British English, and to identify the role morphology plays as a marker of register formality.

Over the last few years, the current authors have engaged in a large-scale project (Laws & Ryder 2014) to identify the characteristics of complex lexemes in spoken language as a function of register variation. This project has involved the analysis of 847 word-initial/word-final morphemes including 575 combining forms, 141 prefixes and 131 suffixes, the last of these consisting of 80 noun-forming, 42 adjective-forming, 5 adverb-forming and 4 verb-forming suffix categories. Together, these form a database of around 1 million tokens. The selection of suffixes was based mainly on the derivational morpheme list reported in Stein (2007), although a few additional suffixes were included from Marchand (1969), Quinion (2002), Dixon (2014), and from consulting the *Oxford English Dictionary* (OED online). The current study focuses on the effect of speech formality on the distributional characteristics of the four principal verb-forming suffixes *-ate*, *-en*, *-ify* and *-ize* in British English. The results of equivalent analyses for complex nouns, adjectives and adverbs, are to be reported elsewhere.

The research questions posed by the current research are:

i. To what extent does speech formality affect suffix category diversity (number of types) and suffix category density (number of tokens) in British English?
ii. To what extent do the register-based vocabulary sets overlap for each verb-forming suffix?
iii. What register-based qualitative differences are conveyed by the complex verb forms in terms of concreteness (physical attributes) and abstractness (cognitive attributes) and the potential of the four suffixes to produce neologisms?

The general prediction, based on the literature on register variation, was that in the more formal context (CG) speakers use a wider repertoire of complex words (greater category diversity), that these derived forms are used more frequently (greater category density), but that the extent of these differences would be considerably smaller than the context effects observed between spoken and written corpora, e.g. Biber et al. (1999) and Schmid (2011). Register-based studies to date have focused on quantitative measures, although some attention to qualitative differences has been provided by Biber et al.

(1999) with respect to the relative concreteness/abstractness of verbal derivatives. The current study explores this approach more systematically by measuring the degree of overlap and non-overlap of complex verb types between corpora, and by examining the extent to which the concreteness/abstractness of verbal derivatives and the production of neologisms vary as a function of context formality.

## 3. Corpus metrics, measures of productivity and morphological categories

The productivity of a morphological category reflects the degree to which "it can be used synchronically in the production of new forms" (Bauer 1983: 18), i.e. the extent to which the overall size of that category continues to increase, rather than remain fixed or decline. Research on the productivity characteristics of derivational categories (Plag et al. 1999; Hay & Baayen 2002, 2003) has employed a number of measures devised originally by Baayen and colleagues (Baayen & Lieber 1991; Baayen 1992, 1993, 1994); the following provides a brief summary of these, but the reader is referred to Baayen (2009) for an overview of these metrics.

Realised Productivity refers to the vocabulary size, or type count, of a derivational category, i.e. its extent of use, or the degree to which the affix has successfully attached to suitable bases. Biber et al. (1999: 400) also refer to this measure as an indicator of productivity, together with the ratio between common (more than 1 token per million) and rare (less than 1 token per million) derivative formations. The type count of each verb-forming suffix provides a snapshot of the vocabulary size that is appropriate for the formality of a particular register. The normalised type count (the number of types per million tokens) represents a 'diversity' measure for that morphological category in relation to a particular register, i.e. the degree of heterogeneity within the vocabulary set as a function of that context. In the current study, normalised type counts, which can be compared directly between the two spoken corpora, will be referred to as 'category diversity'. The token count of that suffix category, on the other hand, provides an indicator of its relevance (repeated use) in that context, in particular with respect to high and low frequency derivatives. In the current study, normalised token counts will be referred to as ‘category density’, and again, this metric permits a direct comparison between different corpora.

In addition to Realised Productivity, Baayen and colleagues (Baayen & Lieber 1991; Baayen 1992, 1993, 1994) have identified two other measures of productivity, Potential and Expanding Productivity, based on the frequency of hapax legomena, i.e. lexemes that occur only once in a corpus. Potential Productivity (*P*) provides a measure of the growth rate of the vocabulary size of the particular morphological category, by dividing the number of hapaxes of a morphological category by the number of tokens of that category in the corpus. If the *P* value of a suffix category is low and the majority of its members have been employed in a corpus of a particular size, it is very unlikely that extending the size of the corpus will lead to the occurrence of more members of that set, i.e. the dataset can be said to be ‘saturated’ (Baayen 2009: 902). If the *P* value is high, greater diversity within the suffix category is likely to result.

Expanding Productivity (*P**) provides a measure of the extent to which the derivational category is contributing to the overall lexical diversity of the corpus, by dividing the number of hapaxes of a morphological category by the number of hapaxes in the corpus. A further derived measure is Global Productivity which is produced by plotting Potential Productivity (*P*) on the *x* axis and Realised Productivity (types) on the *y* axis. Such plots allow the magnitude of these two measures to be considered simultaneously, where greater productivity is associated with larger values on both axes.

The reliance on hapax data in these types of calculations has been questioned due to the difficulty in obtaining reliable comparisons across morpheme categories (Fernández-Domínguez 2013, Säily 2011). Nevertheless, Potential Productivity has been instrumental in successfully predicting the occurrence of neologisms in affix category sets (Baayen 1994) and exploring the relationship between lexeme parsability and the ability of morphological categories to produce neologisms (Hay & Baayen 2002, 2003). Hapaxes and low-frequency items (occurring twice or three times) may fall into one of the following three classes: a lexeme that is in keeping with the register of the corpus; a lexeme that is not consistent with the register of the corpus, such as a rare scientific term; a newly created lexeme, or neologism, which has been coined deliberately to fill a lexical gap. In the analysis reported here, all three of these classes of low-frequency lexemes were considered.

The focus of the current paper is to identify the Realised Productivity (category diversity) and representation (category density) of the four principal verb-forming suffixes in spoken British English depending on speech formality, and the corresponding Potential Productivity of each suffix category. In addition, this paper aims to identify qualitative differences in lexical choice between registers, by considering the nature of high and low frequency derivatives, including neologisms, in order to provide a register-based profile of these suffixes, the specific linguistic characteristics of which are summarised in the following section.

## 4. The function and meaning of verb-forming suffixes in English

The four[1] principal verb-forming suffixes in English, *-ate*, *-en*, *-ify* and *-ize* all have the function of denoting various interpretations of change of state, some of which are shared across the suffix categories, others relate only to a subset. Plag (1999, 2004) has provided a detailed account of the range of semantic interpretations that derivatives formed with these polysemous suffixes can express, a summary of which is presented in Table 1. Since an in-depth discussion of the mapping between semantic category and suffix type is beyond the scope of this paper, the reader is referred to Plag (1999, 2004) for the further information.

As Plag (1999: 126) notes, the process of semantic categorization shown here only represents "regularities" in suffix meaning; individual derivatives may be assigned alternative semantic categories, depending on the context. To adapt an example from Biber et al. (1999: 402), the verb *stabilize* can be paraphrased to mean "become stable" in *inflation stabilized after the election* (inchoative reading), or "make (more) stable", as in *the government stabilized inflation after the election* (causative reading). Therefore, the process of assigning semantic categories to all occurrences of a derivative requires that the actual context in which it occurs be established; this exercise was beyond the scope of this study but is being addressed in a follow-on paper (Laws & Ryder in preparation).

This is a pre-publication version accepted by the International Journal of Corpus Linguistics, to appear in 2018. Please refer to the published version if you wish to quote from it.

**Table 1**. Semantic categories of verb-forming suffixes*

| **Semantic Category** | **Meaning/Paraphrase** | **Examples** *-ate* | *-en* | *-ify* | *-ize* |
|---|---|---|---|---|---|
| Locative | put in(to) X | | | *syllabify, codify* | *hospitalize, containerize* |
| Ornative | provide with X | *chlorinate, nitrogenate* | | *glorify, youthify* | *patinize, texturize* |
| Causative | make (more) X | | *darken, threaten* (transitive) | *diversify, acidify* | *stabilize, oxidize* (transitive) |
| Resultative | make into X | *gelate, activate* | | *teddify, yuppify* | *crystallize, unionize* |
| Inchoative | become X | | *darken, ripen* (intransitive) | *acidify, calcify* | *stabilize, oxidize* (intransitive) |
| Performative | perform X | | | *speechify, boozify* | *philosophize, economize* |
| Similative | act like X | | | *Shelleyfy, Swiftify* | *Powellize, despotize* |

* Adapted from Marchand (1969), Plag (1999) and Plag (2004).

It is immediately clear from Table 1 that the suffixes -*ify* and -*ize* express a far wider range of change-of-state meanings than -*ate* and -*en*. This greater versatility of -*ify* and -*ize* leads to the prediction that derivatives formed with these two suffixes are likely to occur more frequently than the two less productive suffixes (Plag 1999). Marchand (1969) states that only -*ate*, -*ify* and -*ize* constitute the suffixal verb-forming set in English and Bauer (1983: 223) notes that -*en* is only marginally productive. In the next sections, a short summary of the individual characteristics of the four suffixes will be discussed in turn.

### **4.1** The characteristics of -*ate*

The suffix -*ate* originated in Middle English from the Latin past participle -*atus* to anglicize the -*are* infinitive inflection on corresponding Latin verbs (Marchand 1969). However, Marchand (1969: 256) explains that, as a result, the -*ate* ending on such verbs as *imitate* and *terminate* is an "adaptational termination", rather than a derivational suffix. The problems associated with the analysis of verbs bearing the -*ate* suffix have been well-documented by Marchand (1969: 256-8) and Plag (1999: 204-13). Various morphological processes, including backformation from earlier attested forms, account for the presence of numerous non-derivational -*ate* verbs in English. It was only in the 16th century that -*ate* was employed to create verbs from nominal Latinate bases, where no prior Latin verb already existed. Therefore, the adaptational -*ate* ending serves to mark the verbal status of the lexeme (Plag 1999: 211), and the ornative (*nitrogenate*)/resultative (*activate*) meanings are expressed by the derivational suffix (Plag 1999: 205-6, Adams 2001: 22-23).

However, it has not been clear exactly what selection criteria previous researchers have employed to distinguish between adaptational and derivational -*ate*, since frequently examples of the latter provided by some sources have been classified as the former by others. Therefore, the criterion for deciding whether a verb bearing the -*ate* suffix should be included in the current study was to select only those verbs which the OED online analyses explicitly as "base + -*ate*". This conservative procedure may have resulted in a smaller type set, compared with other studies, but it ensured that all -*ate* forms included in the analysis were not derived from backformation or other morphological processes.

Compared with -*ify* and -*ize*, the restricted set of meanings (ornative and resultative) of derivational -*ate*, combined with a number of phonological constraints on the base conditions that permit -*ate* attachment, limit the productivity of this suffix (see Plag 1999 for a detailed discussion). Finally, -*ate* produces de-adjectival and denominal derivatives but the latter constitute the majority (Bauer et al. 2013: 284). In the Biber et al. (1999: 401) study, -*ate* has the second lowest type count of the four verb-forming suffixes across both speech and writing, indicating low productivity and representation in spoken language.

**4.2** The characteristics of -*en*

This is a pre-publication version accepted by the International Journal of Corpus Linguistics, to appear in 2018. Please refer to the published version if you wish to quote from it.

The native suffix *-en* originates from Old English *-nian*, which formed verbs initially from nominal and later adjective stems (Marchand 1969), although the latter category now forms the majority. *-en* attaches to monosyllabic bases ending in an obstruent and, with only a couple of exceptions, these bases are also of native origin (Dixon 2014). As indicated in Table 1, the repertoire of meanings of verbs formed with *-en* is restricted to the following interpretations: causative ("make (more) X") with the transitive form of the verb, and inchoative ("become X") with the intransitive form.

Biber et al. (1999: 401) observe that although a considerable number of verb lexemes have been generated from the *-en* suffix (the second largest type count after *-ize* across the four verb-forming suffixes in conversation), very few of these are rare, indicating that this suffix is only marginally productive, as noted by Bauer (1983: 222).

### **4.3** The characteristics of *-ify*

The change-of-state meaning of the suffix *-ify* is apparent from its Latin root *-ificare* (from *facere*, "make") (Marchand 1969: 300). The attachment of the suffix *-ify* displays a number of phonological restrictions (Dixon 2014: 192-4) and the bases, which are predominantly non-native, can be nouns and adjectives.

Plag (1999: 195-6) notes that *-ify* is less productive than *-ize*, mainly due to phonological restrictions, but the two suffixes share the full repertoire of verb-forming interpretations, as illustrated in Table 1, although the performative and similative meanings are much less common. Plag (2004) illustrates that within the polysemous set of suffix meanings, ornative and locative have reverse interpretations.

Despite the wide range of meanings that the *-ify* suffix conveys, Biber et al.'s (1999: 401) study identifies that, compared with the other three verb-forming suffixes, *-ify* generates the lowest derivative type count in conversation.

This is a pre-publication version accepted by the International Journal of Corpus Linguistics, to appear in 2018. Please refer to the published version if you wish to quote from it.

**4.4** The characteristics of *-ize*

The suffix *-ize* has Greek roots, from transitive and intransitive verbs with the suffix *-ίζειν*; it was later Latinized to *-izare* and the period of greatest productivity was during the Renaissance (Marchand 1969: 318), resulting in predominantly non-native bases which include both nouns and adjectives.

In keeping with the other verb-forming suffixes, *-ize* attachment conditions display a number of phonological restrictions, although these are in complementary distribution with *-ify*. In fact, the two suffixes are intricately related (Bauer et al. 2013: 269) and share the full range of semantic categories illustrated in Table 1 (Plag 1999, 2004).

*-ize* is considered the most productive of the four verb-forming suffixes in English (Bauer 1983: 222, Plag 1999: 122) and this is clearly demonstrated by Biber et al.'s (1999: 401) data where the greatest number of verb types with this suffix occurs in conversation and academic prose (53% and 63%, respectively), and this 10% increase in vocabulary size between these registers is by far the greatest across all four verb-forming suffixes. However, it should be noted that, although the suffixes *-ize* and *-ify* both generate the widest range of different semantic interpretations (see Table 1), these results illustrate that greater polysemy does not necessarily lead to a larger vocabulary size.[2] Biber et al. (1999) conclude that *-ize* is the most productive in producing names of new processes; it can furthermore be surmised that *-ize* is a clear marker of academic register.

## 5. Methodology

In this section, the details of the two spoken registers within the BNC are provided, and the procedure for extracting the complex verb forms for analysis, as well as the analysis procedures employed to compare the two sub-corpora, are described.

**5.1** Data source

The BNC has been used as a frequent data source for investigations on register-related patterns across spoken and written British English from the time of its compilation in the early 1990s and constitutes a useful reference benchmark for cross-study comparisons; register variation studies that used the BNC already cited here include Plag et al. (1999), Cowie (2006), Guz (2009) and Säily (2011, 2016). Since the submission of this paper, a new version of the BNC has been released, the BNC2014 (Love et al. 2017). However, the BNC2014 only contains speech from everyday speech, which is equivalent to the DS component of the original BNC, and therefore was not employed in this study since a comparison of register between everyday speech (DS) and more formal contexts (CG) was the focus of the work reported here.

The objectives for the design of the 10-million word spoken corpus were necessarily different for the two components (Burnard 2007). The DS corpus represents spontaneous conversation from a balanced spread of British English speakers in the UK based on age, gender, social class and geographical region. The CG dataset represents a much wider range of registers in spoken British English that is typical of more formal, technical and public environments; the sub-corpus comprises texts in equal proportion from four context domains: education, business, public/institutional and leisure. Table 2 provides the breakdown of the composition of the two sub-corpora of the spoken BNC.

**Table 2**. Composition of the spoken components of the BNC*

| **Demographically Sampled component (DS)** | **Context-governed component (CG)** |
|---|---|
| Sampled according to: | Categorized by domain: |
| Respondent age | Educational and informative (e.g. educational demonstrations, news commentaries) |
| Respondent sex | Business (e.g. company talks, interviews and sales demonstrations) |
| Respondent social class | Public or institutional (e.g. political speeches, sermons) |
| Geographical region | Leisure (e.g. sports commentaries, club meetings, chat shows broadcast on television or radio) |
| **Total DS token count:** 4,233,962 words | **Total CG token count:** 6,175,896 words |

* Adapted from Hoffman et al. (2008: 34).

These two sub-corpora were employed in the present study to establish the distributional characteristics of verb-forming suffixes, as a function of speech formality in British English.

**5.2** Data preparation and analysis

The extraction of complex words from the DS and CG subcomponents of the spoken BNC was conducted using *BNCweb* (Hoffman et al. 2008). Raw data files were extracted for each search string (*suffix), together with all possible inflections for verbs: third person singular present -*s*, present participle -*ing*, past simple and past participle -*ed*. The token frequency and Part of Speech (PoS) assignment was recorded for each entry.

The grammar tagger employed by *BNCweb* is CLAWS C5 tagset (Garside & Smith 1997); it is reported on the website that the overall tagging error rate for the spoken corpus is 1.17% and that the PoS ambiguity rate is 3.00%. Ambiguous tagging occurs when the grammar tagger is unable to determine whether a word is, for example, a noun, an adjective or a verb, as in *standardizing*. Ambiguous tags are very likely to occur with spoken language which is typically fragmented; in these cases a paired ambiguous tag, e.g. VVG-NN1 (present participle or singular noun) is provided in place of a single PoS tag. In the current study, all ambiguities of this type were resolved by checking the context of each word in the BNC. Where no ambiguity was flagged, but the word class assigned seemed unusual or unlikely, the PoS was also checked in the BNC; for example *summarise* was assigned the PoS NN1, singular noun by the grammar tagger, whereas, on inspection of the context, it was found, unsurprisingly, to be VVB, the finite base form of the verb, and was therefore recorded accordingly in the dataset with the correct PoS.

The raw word lists were then processed to produce the dataset of complex words by eliminating simplex words and checking potential complex word candidates against the OED online; the first criterion for inclusion of a complex word was that the etymological information used the formulation “base + suffix”. Proper nouns formed from derivatives were excluded from the dataset. Complex words with multiple derivational suffixes were included in the dataset based on the outermost suffix so, for example, the item *industrialize* was allocated to the verb-forming -*ize* suffix category, but not to a set for the adjective-forming suffix -*al*. The final dataset used for the analysis consisted of normalised type and token frequencies for each of the four verb-forming suffixes for the DS and CG sub-corpora.

As mentioned in Section 3, normalised type and token counts were employed as measures of category diversity and category density. It is customary to use normalised token counts to compare the relative occurrence of target lexemes between corpora (McEnery & Hardie 2012: 49-50), since the procedure requires target token frequencies to be divided by the total number of tokens in the respective corpora. However, the comparison of type frequencies across corpora is not so straightforward because it is not feasible to arrive at a normalised type count based on a denominator that represents the total number of types in a corpus. In order to overcome this problem it was decided to calculate normalised type counts using the total number of tokens in the respective corpora as the denominator (following the normalised token count procedure). The inevitable drawback of this approach is that with large corpora, such as the DS and CG, normalised type count values are underestimated because of the non-linear relationship between type and token growth rate: as the corpus size increases, type frequency growth rate slows down (Baayen 2008: 222-4), therefore the larger the corpus, the more deflated the normalised type frequency becomes. It is therefore recognised that the normalised type counts reported in this study are conservative.

The main statistical analysis procedure employed here for token count comparisons was the log-likelihood test (LL), since it is preferred for larger corpora and makes no assumptions of normality with respect to the distribution of data (Dunning 1993). The LL test relies on the normalisation procedure based on total corpus token count and is therefore ideal for testing the significance of differences between token counts of target lexemes. In the current study, the LL test was also used for comparing type frequencies but, for the reasons stated above, these values were necessarily deflated through normalisation based on total token count; therefore the likelihood of some type count comparisons reaching statistical significance was reduced, resulting in a conservative assessment of type differences across corpora. The Bayesian Information Criterion (BIC), which is recommended for corpus analyses, is also reported for each

LL value in order that the effect size of the statistic could be evaluated: the criterion value was set to 6 (Wilson 2013: 6).

One of the objectives of the study was to compare type and token differences of the four verb-forming suffixes both within and between the DS and CG corpora, thus employing the datasets twice. Such multiple comparisons can lead to Type I errors, since the likelihood of obtaining significant outcomes may be inflated. To overcome this, the Bonferroni correction was applied: where multiple pairs were compared, the standard $\alpha$ of 0.05 was divided by 2, yielding a corrected $\alpha$ of 0.025. However, the minimum significance threshold for $p$ was set to 0.01 to provide an even more conservative cut-off point for statistical significance. In fact, all significant multiple comparisons reported in Section 6 reached significance at the 0.01 level or above.

## 6. Results and discussion

This section is divided into four parts. Firstly, the overall distribution patterns of the four verb-forming suffixes are reported. This is followed by the analysis of the effect of spoken register on each suffix category. The third section addresses the degree to which register-based vocabulary sets overlap for each verb-forming suffix, and the final section examines the qualitative differences between register-based vocabulary sets in terms of concreteness/abstractness and the nature of neologisms, as a function of register formality.

### **6.1** Distribution patterns of *-ate*, *-en*, *-ify* and *-ize*

Before examining the effect of register, the distributional characteristics of the four verb-forming suffixes within the context of the BNC spoken corpus are compared. Table 3 presents the raw and normalised (occurrences per million) type and token frequencies of the suffixes *-ize*, *-en*, *-ify* and *-ate* in type rank order. In accordance with expectations, and in line with Biber et al.'s (1999) data, the *-ize* category was overwhelmingly the most diverse (greatest type frequency) and densely represented (greatest token frequency); these values significantly exceeded those of the second ranked suffix *-en* (Types: LL=58.52, *df*=1, *p*<0.0001, BIC=41.67; Tokens: LL=2,798, *df*=1, *p*<0.0001, BIC=2,781.23), which also ranked second in Biber et al.'s (1999) data. The type count differences between *-en* and *-ify* (59 and 54) failed to reach significance, but the token count for *-ify* was significantly greater than that for *-en* (LL=201.68, *df*=1, *p*<0.0001, BIC=184.82). The diversity and density of verbs formed with *-ate* were the lowest of the four suffixes and differences with adjacent ranked values reached statistical significance (Types: *-ify* vs. *-ate*, LL=11.84, *df*=1, *p*<0.001, although the size effect did not reach criterion; Tokens: *-en* vs. *-ate*, LL=486.73, *df*=1, *p*<0.0001, BIC=469.88). In Biber et al.'s (1999) data, although type count differences between *-ify* and *-ate* are close, *-ify* represented the smallest category diversity in that study.

**Table 3**. Comparison of raw and normalised type, token and hapax frequencies of the 4 verb-forming suffixes

| | | **Raw Frequencies** | | | **Frequencies per million** | | |
|---|---|---|---|---|---|---|---|
| **Suffix** | **Example** | **Types** | **Tokens** | **Hapaxes** | **Types** | **Tokens** | **Hapaxes** |
| *-ize* | *criticize* | 173 | 5,062 | 48 | 17 | 486 | 4.61 |
| *-en* | *frighten* | 59 | 1,082 | 7 | 6 | 104 | 0.67 |
| *-ify* | *classify* | 54 | 1,846 | 11 | 5 | 177 | 1.06 |
| *-ate* | *activate* | 24 | 290 | 9 | 2 | 28 | 0.86 |
| | **Totals** | 310 | 8,280 | 75 | 30 | 795 | 7.20 |

This initial analysis reveals that the *-ize* morphological category makes up 56% of the total type count of verbal derivatives, which is very close to the approximate 53% observed in Biber et al.'s (1999) data (see Section 4.4); the lowest contributor is the *-ate* category with just 8%. *-en* and *-ify* both contribute around 18% each, but *-ify* is significantly better represented in the corpus in terms of tokens than *-en*; this finding is slightly at odds with Biber et al.'s (1999) data where *-ify* only contributes around 8% to the total speech type count for verb derivatives. In the next section, the effect of spoken register on the relative contribution of these suffix categories will be examined.

**6.2** Category diversity and density between DS and CG registers

The corpus-specific profiles (DS and CG) of the normalised type and token frequencies of the suffixes *-ize*, *-en*, *-ify* and *-ate* are rank-ordered by type count in Table 4. The columns labelled "Types CG/DS" and "Tokens CG/DS" provide a measure of the magnitude of the difference in type and token frequencies as a function of register; these have been calculated by dividing the normalised CG Type or Token values by the equivalent normalised DS values to reflect the factor by which the CG counts exceed those observed in the DS corpus.

**Table 4**. Normalised type and token frequencies of the verb-forming suffixes between the DS and CG corpora*

| | | **Types / million** | | | **Types** | **Tokens/million** | | | **Tokens** |
|---|---|---|---|---|---|---|---|---|---|
| **Suffix** | **Example** | **DS** | **CG** | **Sig** | **CG/DS** | **DS** | **CG** | **Sig** | **CG /DS** |
| *-ize* | *criticize* | 16 | 27 | *** | 1.71 | 263 | 640 | **** | 2.44 |
| *-en* | *frighten* | 11 | 9 | NS | 0.80 | 81 | 119 | **** | 1.14 |
| *-ify* | *classify* | 8 | 8 | NS | 0.94 | 35 | 275 | **** | 7.87 |
| *-ate* | *activate* | 2 | 4 | NS | 1.68 | 6 | 43 | **** | 7.60 |
| | **Totals** | 37 | 47 | * | 1.27 | 385 | 1,077 | **** | 2.80 |

* The shaded cells indicate the significantly larger of the two values compared. *$p$<0.05; ***$p$<0.001; ****$p$<0.0001.

The difference between the normalised totals of complex verbs types occurring in the two sub-corpora (37 and 47) only just reached significance, although the size effect did not reach criterion (LL=5.98, $df$=1, $p$<0.05), indicating that a greater variety of verb lexemes was observed in the more formal CG corpus. The overall significant difference in type counts was attributable principally to the *-ize* verb-forming suffix (LL=14.68, $df$=1, $p$<0.001, even though the size effect did not reach criterion), where category diversity for the CG corpus was 1.71 times greater than it was for the DS corpus (Table 4). This result provides further empirical support to Biber et al.'s (1999) finding that *-ize* has the greatest Realised Productivity of the verb-forming suffixes in conversation and that as the register becomes more formal (from conversation to the written context in Biber et al.'s (1999) case) the diversity of *-ize* derivatives increases. This result is contrasted with the other three suffixes, where type counts failed to differ significantly between the two spoken registers. However, it must be noted that although *-ate* type counts increased markedly between the DS and CG corpora, this difference failed to reach significance; further similarities between *-ate* and *-ize* reoccur in this study.

In terms of token frequencies, the differences between the two corpora were more marked (Table 4). The overall density of complex verbs in the CG corpus was 2.80 times greater than that of the more colloquial DS context; total token differences were highly significant (LL=1,666.72, $df$=1, $p$<0.0001, BIC=1,650.56). Here, all four suffixes contributed to this overall register effect, producing statistically robust token differences ($p$<0.0001 in each case), the greatest of which occurred with complex verb forms ending in *-ify* (LL=1,008.56, BIC=992.40), followed by *-ize* (LL=795.07, BIC=778.91), *-ate* (LL=155.37, BIC=139.22) and lastly *-en* (LL=36.41, BIC=20.25). The observation that the suffix *-ify* produced a greater register effect than *-ize* is rather unexpected, given the greater category diversity of the former and the fact that these two suffixes share a common repertoire of meanings (see Table 1); nevertheless, it can be concluded that the comparatively smaller *-ify* type set is better represented (denser) in more formal settings (7.87 times more frequently in the CG than the DS context), whereas each item in the considerably richer *-ize* type set is reused comparatively less often in the more formal context (2.44 times more frequently in the CG than the DS context). The suffix *-ate* produced the second highest CG/DS ratio (7.60) indicating that, like *-ify*, a restricted vocabulary set was represented many times more frequently in the CG than the DS corpus. By contrast, verbs formed with *-en* did not increase markedly between the DS and CG corpora (CG/DS ratio =1.14), suggesting that only a restricted set of verb types occur more frequently in the more formal register.

As mentioned in Section 3, hapax legomena provide a means for determining probabilistic measures of productivity (Baayen 2009); they represent the "used-once" members of a category in a particular vocabulary set. They may include neologisms, very low frequency types, or types that are rarely used in the particular context. The normalised hapax values for each corpus in Table 5 ("Hapaxes/million") correspond to Baayen's (2009) notion of relative Expanding Productivity (see Section 3), i.e. the contribution of each category to the overall vocabulary set.

This is a pre-publication version accepted by the International Journal of Corpus Linguistics, to appear in 2018. Please refer to the published version if you wish to quote from it.

**Table 5**. Normalised hapax frequencies and Potential Probability values of the verb-forming suffixes between the DS and CG corpora*

| | | **Hapaxes/million** | | | **Potential Productivity *P*** | | |
|---|---|---|---|---|---|---|---|
| **Suffix** | **Example** | **DS** | **CG** | **Sig** | ***P*(DS)** | ***P*(CG)** | ***P*(DS)/*P*(CG)** |
| *-ize* | *criticize* | 6.14 | 7.93 | NS | 0.023 | 0.012 | 1.92 |
| *-en* | *frighten* | 4.02 | 0.81 | *** | 0.049 | 0.007 | 7.00 |
| *-ify* | *classify* | 4.02 | 1.33 | ** | 0.115 | 0.004 | 28.75 |
| *-ate* | *activate* | 0.94 | 1.30 | NS | 0.167 | 0.030 | 5.57 |
| | **Totals** | 14.27 | 11.37 | NS | | | |

* The shaded cells indicate the significantly larger of the two values compared. ***p*<0.01; ****p*<0.001.

The distribution of hapaxes follows that of the corresponding normalised types in Table 4, i.e. as type count reduces from *-ize* to *-ate*, so does the hapax count. The density of *-ize* and *-ate* hapaxes does not differ significantly between the two corpora, whereas significantly fewer hapaxes were identified for *-en* (LL=12.23, *df*=1, *p*<0.001) and *-ify* (LL=8.92, *df*=1, *p*<0.01) in the CG corpus, although the size effect criterion was not met for these comparisons. These results indicate that the contribution of complex verb types bearing *-ize* and *-ate* remain similar across the corpora, with a slightly higher contribution in the CG. By contrast, the complex verb types bearing *-en* and *-ify* are not expanding as the context becomes more formal, instead the two vocabulary sets are reused more frequently. This conclusion is also reflected in the Potential Productivity measures in Table 5, where the small *P* values for *-en* and *-ify* in the CG corpus indicate a low probability of new members entering the category set compared with the DS; in the case of *-ify* the chances are as much as 28.75 times lower, as indicated by the "*P*(DS)/*P*(CG)" measure, which represents the factor by which the *P* value for DS exceeds that for CG.

To explore these register effects further, Figure 1 shows the balance between low (less than one token per million) and high frequency types (more than one token per million) across the two corpora. The following analysis extends the examination of Potential Probability above beyond hapax counts, by considering all low frequency types.

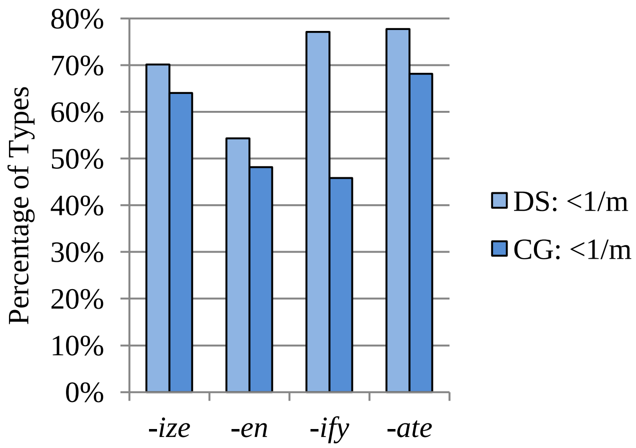


**Figure 1**. Percentage of types with token frequencies less than 1 per million by suffix category

It is clear from Figure 1 that the main difference between the two corpora, in terms of token frequencies, concerns the *-ify* suffix category, but first we will consider the relative distributions of the other three categories. It will be recalled from Table 4 that the relative vocabulary set for *-ize* increases significantly between the DS and CG corpora and Table 5 demonstrates that this category is expanding to about the same extend in both corpora. Figure 1 illustrates that the percentage of low frequency types decreases slightly between registers (a 6% drop from 70% to 64%). This result aligns with Biber et al.'s (1999) findings, where the percentages were roughly as follows: 66% for conversation and 50% for academic prose (Biber 1999: 401). Therefore, as formality increases, low frequency lexemes are, unsurprisingly, used more frequently.

Although category diversity does not differ between DS and CG for *-en* and *-ate*, again the proportion of low frequency types drops as formality increases (a 6% drop from 54% to 48% for *-en* and a 10% drop from 78% to 68% for *-ate*). Biber's (1999) figures show a more dramatic change between conversation and academic prose from roughly 66% to 25% for *-en* and from roughly 100% to 25% for *-ate*; these figures suggest that when a more extreme comparison is made (i.e. between conversation and writing, as opposed to between two spoken registers), the reduction in low frequency members of these two suffix categories is emphasised. Turning now to the equivalent percentages for *-ify*, it was observed in Table 5 that this suffix category demonstrated extremely low potential growth in the CG corpus compared to the DS and, in Figure 1, a very marked decrease in the proportion of low frequency types was observed between the corpora (a 31% drop from 77% in the DS to 46% for the CG) which, while being very much in keeping with the scale of the decrease reported in Biber's (1999) study (from roughly 55% in conversation to roughly 25% for academic prose) is, however, substantially greater than the decreases observed for the other three suffixes in Figure 1.

Taken together, the findings in this section illustrate that as speech register formality increases, category diversity (number of types per million tokens) of verbal derivatives does not increase noticeably, with the exception of the *-ize* category, where the vocabulary set does increase significantly between the DS and CG corpora. Thus, degree of speech formality directly affects the type count of verbal derivatives in the *-ize* suffix category. With respect to token counts, the proportion of low frequency complex verb types decreases as register formality increases; or conversely, the need to reuse rarer, more specialised members of a category set increases. Such compositional differences were observable between the DS and CG corpora, thus providing very clear evidence that register formality affects representation of derivatives in speech, in particular in relation to complex verbs in the *-ify* derivative category.

These results are explored further in the next section with respect to the degree to which verbal derivatives are shared by the two speech registers and to what extent they are unique to the setting.

### **6.3** The overlap between register-based vocabulary sets

The subsequent analysis was designed to identify the proportion of complex verb types that (i) were shared by the two corpora, (ii) only occurred in the DS corpus, and (iii) were unique to the CG corpus. The relative proportions of these three categories are presented in Figure 2 for the DS and CG corpora.

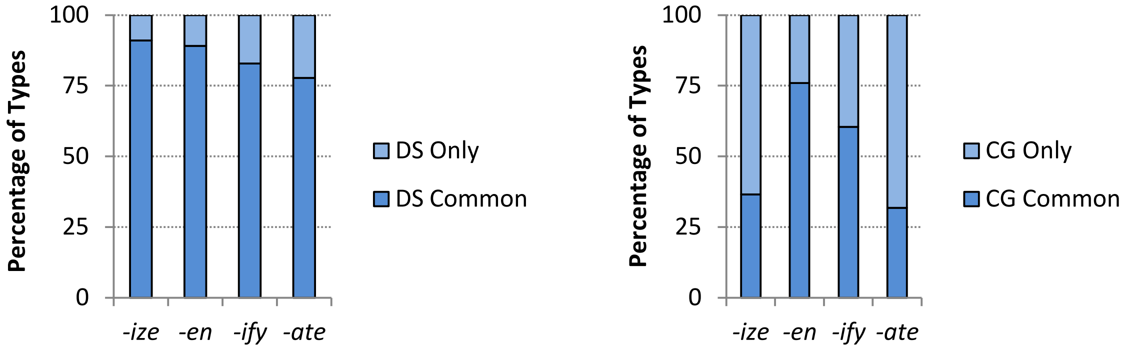


**Figure 2**. Percentage of types commonly shared and unique to DS and CG corpora per affix

Figure 2a illustrates that, regardless of suffix category, the majority of derivative types occurring in the DS corpus also occurred in the CG corpus; only a small percentage of the DS vocabulary set for each morphological category was unique to the less formal context. In contrast, Figure 2b shows a very different pattern emerging across the suffix categories for the CG corpus. Here, the majority of *-en* types (76%) also occur in the DS corpus, suggesting that the majority of the vocabulary set is unaffected by speech register. In Section 4.2, it was noted that the *-en* suffix is native in origin, as are the majority of the bases it attaches to. As a result, *-en* derivatives are less likely to be specialised in nature; it is therefore not surprising that complex verbs in this category may be appropriate for both more and less formal contexts. This proposition is explored further in Section 6.4.

It is also illustrated in Figure 2b that verb types formed with *-ify* are less specialised than those ending in *-ize* and *-ate*, since a relatively large proportion (60%) also occur in the DS corpus, with relatively fewer being unique to the CG corpus. However, *-ify* is a Latinate suffix and therefore more likely to be associated with scientific or technical terms, therefore it is less expected that verbal derivatives in this category should be common to the two speech registers. By contrast, in accordance with expectations relating to the association of Latinate derivatives and more specialist terms, the *-ize* and *-ate* categories show the highest proportion of unique forms in the more formal CG corpus (63% and 68%, respectively). These findings taken together indicate that, although the actual type counts for these two verb-forming suffixes represent the smallest and the largest across the four suffixes analysed, lexical choice is strongly affected by speech register differences.

To summarise, the choice of *-en* verbal derivatives is least affected by speech register formality, whereas the suffix categories *-ize* and *-ate* display considerable heterogeneity in more specialised speech environments. The next section examines lexical choice differences across registers.

**6.4** Qualitative differences in register-based vocabulary sets

The results so far indicate that the vocabulary overlap between the two speech registers is greatest for *-en* and least for *-ize* and *-ate*. Appendix 1 presents the complex verb types that are common to the two corpora, grouped according to the size of the token increase (or decrease) between the DS and CG datasets, from not statistically significant to significant at the 0.0001 level. As also observed by Biber et al. (1999: 402), it will be noted here that as the difference in token density increases between the corpora, the more abstract the lexical items become for the Latinate forms *-ize*, *-ify* and *-ate* (e.g. concrete terms such as *pulverize*, *magnify* and *dehydrate* occur in the common set that does not increase significantly in token frequency between the corpora, but more shared abstract terms, such as *summarize*, *identify* and *evaluate* increase very significantly). This is not so marked for the native suffix category *-en*, the shared derivatives of which tend to refer to more physical change-of-state processes (e.g. *strengthen*, *threaten* and *widen*), even when their token frequencies increase significantly between registers. It is also the case that four of the *-en* derivatives (underlined in Appendix 1), occurred significantly more frequently in the DS than CG corpus.

Inspection of the verbal derivatives that occurred with a frequency of over 1 per million in the CG corpus (Appendix 2), reveals a similar pattern. In the more formal context, the Latinate categories *-ize*, *-ify* and *-ate* contain abstract, cognitive terms such as *rationalize*, *exemplify* and *formulate*; whereas the *-en* category contains more concrete, physical terms as *awaken*, *darken*, *quicken* and *toughen*.

These findings endorse the earlier conclusions that the *-en* verbal category is least affected by speech register differences, both in terms of the diversity of the category set and its representation, as these derivatives denote more concrete, physical, non-specialised processes and are appropriate for a variety of speech contexts.

However, despite the more predictable nature of this suffix category, one neologism was identified in the DS corpus, *hotten*, indicating some remnants of Potential Productivity:

(1) […] it doesn’t work like that it takes about [pause] cos it’s cold it’s gonna take about an hour [pause] to warm up [pause] and that won’t come on [pause] that’ll just cu-- keep coming on and off all the time [pause] till it’s *hottened* up. (KBF 3237)

As noted in Section 4.2, the suffix *-en* is considered only marginally productive (Bauer 1983: 222), yet in Table 5, its Potential Productivity value in the DS is higher than that for *-ize*, indicating potential for generating neologisms (Baayen 1994). The appropriateness of the term *hotten* in the context of Example (1) demonstrates that even marginally productive affix categories, such as *-en*, continue to be productive, albeit to a limited extent; an observation noted by Baayen (2009) in relation to the neologism *coolth*, derived from the unproductive affix category *-th* (Bauer 1983: 49).

At the opposite end of the continuum, *-ize* derivatives are highly characteristic of specialised language as shown by the significant difference in category size between the DS and CG corpora and the large proportion of forms that occur uniquely in the CG corpus (Figure 2b); some examples of the most frequent of which are listed at Appendix 2. Following Biber et al. (1999), more frequent is defined here as more than 1 token per million. Furthermore, the most numerous examples of neologisms identified were formed with the *-ize* suffix: *assassinize*, *corpusize*, *panelize* and *sanctionalize*, as seen in Examples (2) to (5), all of which occurred in the CG corpus.

(2) I’m a very against character assassination erm I don’t think that Brenda was character *assassinizing*. (D91 335)

(3) I mean, would you assume Jeremy we’re going to be *Corpusized*? (KRY 4) Yeah, you’re going to be *corpusized*, yes. (KRY 5)

(4) […] Russell by this work became […] the main founder of this kind of logic which by the much more sophisticated symbolic apparatus erm is able to *panelize* a much wider range of logical phenomena, [...] (KS3 144)

(5) “Well [pause] in effect it says that how that it’s now the Party is *sanctionalizing* absolute egalitarianism, the aim that landlords, K M T officers, everyone’s going to get equal distribution of land. (JJL 450)

The morphological category that behaves in a similar fashion to *-ize* is *-ate*, although its category size is the smallest of the four and no neologisms were identified. The characteristics of the *-ify* category, on the other hand, seem to fall somewhere between *-en* and the *-ize*/*-ate* pair: its category size is quite limited, indicating restricted diversity, but the members that are common to both registers increase most markedly in token frequency as formality increases (Figure 2b and Appendix 1), indicating that these derivatives are employed more appropriately in a more specialised environment. Again, no neologisms were identified in this suffix category.

## 7. Conclusion

This research has cast new light on the nature of register effects on the diversity and relevance of verbal derivatives in spoken British English. It has demonstrated that the use of verbal derivatives increases register formality. Differences in register variation between the DS and CG components of the spoken BNC were found along a number of dimensions and illustrate that, by separating the components of the corpus, both quantitative and qualitative characteristics of the four verb-forming derivatives in speech are revealed. Despite the fact that *-ize*, *-en*, *-ify* and *-ate* all have a causative function with shared interpretations, these suffix categories each possesses an individual profile with respect to diversity and representation across speech registers.

In line with the findings of Biber et al. (1999), Plag (1999) and Schmid (2011), this study has provided additional empirical evidence that the verbal category *-ize* displays the greatest Realised Productivity (attested category diversity), representation (category density) and Potential Productivity (potential to create neologisms). In addition to these established characteristics of *-ize*, this study also revealed that, although category diversity increased significantly as context formality increased, the representation of these derived forms did not rise as markedly as verbal suffixes with more restricted productivity, such as *-ify* and *-ate*.

Far less attention has been given to the other three verb-forming suffixes and the research reported here revealed that the type and token frequencies of the native suffix category *-en* were least affected by context formality, whereas register-based differences were most apparent with respect to the three Latinate suffix categories *-ize*, *-ify* and *-ate*, as would be expected given the greater likelihood of derivatives of classical origins to occur in more formal contexts. However, no relationship was found between the range of meaning interpretations afforded by each suffix category and category diversity, and the representation patterns of these suffixes in different contexts were not predictable from the size of their respective vocabulary sets.

These results have a direct impact on the design of experimental studies where word frequency is one of the variables; the complexity of the construct of frequency and the multitude of factors that affect it has been demonstrated by Baayen et al. (2016). Different profiles of exposure to English of participants, ranging from children at the various Key Stages 1-5, or language learners of English at various levels of proficiency, will reflect different vocabulary knowledge characteristics. Therefore, this study indicates that the process of controlling word frequency for experimental purposes requires that appropriate register-based values be taken into consideration. The findings reported here illustrate that using the BNC as a single spoken corpus disguises important underlying differences between the DS and CG components and that the separation of whole spoken corpus into its two respective components of everyday spoken British English and the variety of British English typically used in broadcasts, news commentaries and company presentations, provides a useful method for providing the norms to be employed in such studies.

## Acknowledgements

We are grateful to the British Academy and Leverhulme Trust for the grant SG150770 awarded to Jacqueline Laws for the purposes of funding the extraction of complex words from the two subcomponents of the British National Corpus.

## Notes

**1.** The Old English suffix *-le*, as in *scuttle* and *crackle*, denotes the short and repetitive features of movements and sounds. Based on the etymological information provided by Marchand (1969), this suffix has not been included in the verb-forming set analysed here for the following reasons: (i) where there is a discernible base, it is often a verb; therefore this suffix cannot be classified as verb-forming; (ii) many *-le* derivatives predate their bases, as in *twinkle* and *fizzle*, and therefore the derivational status of the suffix is debatable; in fact Marchand (1969: 323) suggests that *-le* “is not a derivative suffix proper from existing roots”; (iii) several *-le* verbs do not have identifiable bases, as in *ramble* and *whistle*. Finally, the *-le* suffix has a process rather than a change-of-state meaning and therefore from a semantic perspective does not belong to the class of verb-forming suffixes analysed in the current study.

**2.** The nature of the competition between the polysemous suffixes *-ify* and *-ize* is discussed in detail by Plag (1999) and is outside the scope of this study.

## References

Adams, V. (2001). *Complex Words in English*. Harlow: Pearson Education Limited.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 109–149). Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1992* (pp. 181–208). Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (1994). Productivity in language production. *Language and Cognitive Processes*, *9*(3), 447–469.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* Cambridge: Cambridge University Press.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 899–919). Berlin: Mouton de Gruyter.

Baayen, R. H., & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, *29*(5), 801–843.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174-1220.

Bauer, L. (1983). *English Word-Formation*. Cambridge: Cambridge University Press.

Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Burnard, L. (2007). *Reference Guide for the British National Corpus*. Oxford: Oxford University Computing Services. Retrieved from http://www.natcorp.ox.ac.uk/docs/URG/ (last accessed November 2016).

Cowie, C. (2006). Economical with the truth: Register categories and the functions of *-wise* viewpoint adverbs in the British National Corpus. *ICAME Journal*, *30*, 5-36.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*, 61-74.

Dixon, R.M.W. (2014). *Making New Words: Morphological Derivation in English*. Oxford: Oxford University Press.

Fernández-Domínguez, J. (2013). Morphological productivity measurement: Exploring qualitative versus quantitative approaches. *English Studies*, *94*(4), 422–447.

Friginal, E, Pearson, P., Di Ferrante, L., Pickering, L., & Bruce, C. (2013). Linguistic characteristics of AAC discourse in the workplace. *Discourse Studies, 15*(3) 279–298.

Garside, R., and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G., and McEnery, A. (Eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102-121), London: Longman.

Guz, W. (2009). English affixal nominalizations across language registers. *Poznań Studies in Contemporary Linguistics*, *45*(4), pp. 447–471.

Hay, J., & Baayen, R. H. (2002). Parsing and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 2001*, (pp. 203–235). Dordrecht: Kluwer Academic Publishers.

Hay, J., & Baayen, R. H. (2003). Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, *15*(1), 99–130.

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund Prytz, Y. (2008). *Corpus Linguistics with BNCweb –APractical Guide*. Frankfurt am Main: Peter Lang.

Laws, J. & Ryder, C. (2014). Getting the measure of derivational morphology in adult speech a corpus analysis using *MorphoQuantics*. *Language Studies Working Papers*, *6*, 3-17.

Laws, J., & Ryder, C. (In preparation). Verb-forming suffixation: Semantic category distributions as a function of register.

Love, R., Dembry, C., Hardie, A., Brezina V., & McEnery T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. Special Issue of *International Journal of Corpus Linguistics*, *22*(3).

Marchand, H. (1969). *The Categories and Types of Present-day English Word-formation: A Synchronic-diachronic Approach* (2nd ed.). Munich: C. H. Beck'sche Verlagsbuchhandlung.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing, 29,* 3-15.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Plag, I. (1999). *Morphological Productivity: Structural Constraints in English*. Berlin: Mouton de Gruyter.

Plag, I. (2004). Syntactic category information and the semantics of derivational morphological rules. *Folia Linguistica, 38*(3-4), 193-225.

Plag, I., Dalton-Puffer, C., & Baayen, R. H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics*, *3*(2), 209-228.

Quinion, M. (2002). *Ologies and Isms: Word Beginnings and Endings*. Oxford: Oxford University Press.

Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, *7*(1), 119-141.

Säily, T., & Suomela, J. (2017). *Types2*: Exploring word-frequency differences in corpora. In T. Hiltunen, J. McVeigh, & T. Säily (Eds.) *Big and Rich Data in English Corpus*

*Linguistics: Methods and Explorations. Studies in Variation, Contacts and Change in English*. Helsinki: VARIENG.

Schmid, H-J. (2011). *English Morphology and Word-formation: An Introduction*. Berlin: Erich Schmidt Verlag.

Stein, G. (2007). *A Dictionary of English Affixes: Their Function and Meaning*. Munich: Lincom Europa.

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger & A. Koll-Stobbe (Eds.) *New Approaches to the Study of Linguistic Variability. Language Competence and Language Awareness in Europe* (pp. 3-11). Frankfurt: Peter Lang.

**Appendix 1.** Significance level of token differences of complex verbs shared by the DS and CG corpora, where the CG token count >DS token count (DS > CG for underlined items)

| | | |
|---|---|---|
| | NS between DS and CG | *acclimatize*, *anaesthetize*, *baptize*, *cannibalize*, *colonize*, *epitomize*, *familiarize*, *fantasize*, *hypnotize*, *institutionalize*, *itemize*, *materialize*, *memorize*, *modernize*, *neutralize*, *ostracize*, *pasteurize*, *patronize*, *practise*, *pulverize*, *realize*, *scrutinize*, *socialize*, *standardize*, *sterilize*, *synchronize*, *terrorize*, *vandalize*, *vaporize*, *visualize* |
| ***-ize*** | LL> 3.84, $df$=1, $p$<0.05 | *apologize*, *crystallize*, *economize*, *fertilize*, *jeopardize*, *nationalize*, *penalize*, *pressurize*, *privatize*, *revitalize*, *stabilize* |
| | LL> 6.63, $df$=1, $p$<0.01 | *formalize*, *specialize*, *sympathize* |
| | LL> 10.83, $df$=1, $p$<0.001 | *categorize*, *finalize*, *victimize* |
| | LL> 15.13, $df$=1, $p$<0.0001 | *authorize*, *criticize*, *emphasize*, *equalize*, *generalize*, *maximize*, *minimize*, *mobilize*, *organize*, *publicize*, *recognize*, *subsidize*, *summarize*, *utilize* |
| | NS between DS and CG | *blacken*, *brighten*, *christen*, *dampen*, *deafen*, *deepen*, *enlighten*, *flatten*, *glisten*, *harden*, *hearten*, *lighten*, *liven*, *loosen*, *moisten*, *quieten*, *ripen*, *sharpen*, *sicken*, *slacken*, *smarten*, *stiffen*, *straighten*, *sweeten*, *thicken*, *tighten*, *waken* |
| ***-en*** | LL> 3.84, $df$=1, $p$<0.05 | *fasten*, *frighten*, heighten, lessen, *shorten*, *soften* |
| | LL> 6.63, $df$=1, $p$<0.01 | *broaden*, *hasten*, *liken*, *weaken*, *worsen* |
| | LL> 10.83, $df$=1, $p$<0.001 | |
| | LL> 15.13, $df$=1, $p$<0.0001 | *strengthen*, *threaten*, *widen* |
| ***-ify*** | NS between DS and CG | *electrify*, *fortify*, *magnify*, *mortify*, *purify*, *sanctify*, *signify*, *terrify*, *verify* |

This is a pre-publication version accepted by the International Journal of Corpus Linguistics, to appear in 2018. Please refer to the published version if you wish to quote from it.

| | | |
|---|---|---|
| | LL> 3.84, *df*=1, *p*<0.05 | *certify*, *glorify*, *intensify*, *rectify*, *testify*, *unify* |
| | LL> 6.63, *df*=1, *p*<0.01 | *classify*, *diversify* |
| | LL> 10.83, *df*=1, *p*<0.001 | *ratify*, *simplify* |
| | LL> 15.13, *df*=1, *p*<0.0001 | *clarify*, *crucify*, *identify*, *justify*, *modify*, *notify*, *qualify*, *quantify*, *satisfy*, *specify* |
| | NS between DS and CG | *dehydrate*, *incapacitate*, *insulate*, *orchestrate* |
| | LL> 3.84, *df*=1, *p*<0.05 | *motivate* |
| ***-ate*** | LL> 6.63, *df*=1, *p*<0.01 | |
| | LL> 10.83, *df*=1, *p*<0.001 | |
| | LL> 15.13, *df*=1, *p*<0.0001 | *activate*, *evaluate* |

**Appendix 2.** Complex verbs occurring in the CG corpus only with high frequency (1 token per million or more, following Biber et al. 1999)

| | **CG corpus only** |
|---|---|
| ***-ize*** | *antagonize*, *capitalize*, *centralize*, *characterize*, *civilianize*, *conceptualize*, *criminalize*, *industrialize*, *legalize*, *mechanize*, *moisturize*, *optimize*, *palletize*, *personalize*, *polarize*, *prioritize*, *rationalize*, *symbolize* |
| ***-en*** | *awaken*, *darken*, *quicken*, *sadden*, *toughen* |
| ***-ify*** | *amplify*, *exemplify*, *falsify*, *gratify*, *liquefy* |
| ***-ate*** | *facilitate*, *formulate*, *orientate* |

*Authors' addresses*

Jacqueline Laws

Department of English Language & Applied Linguistics

University of Reading

Whiteknights

Reading, RG6 6AW

United Kingdom

j.v.laws@reading.ac.uk

Chris Ryder

Department of English Language and Applied Linguistics

University of Reading

Whiteknights

Reading, RG6 6AW

United Kingdom

c.ryder@pgr.reading.ac.uk